Revision of the agriculture production data domain in FAOSTAT

7 December 2016

The FAO Statistics Division is undertaking a methodological overhaul of its main statistical domains to improve the quality of the data disseminated through FAOSTAT. The Agricultural Production domain is the first one for which the data disseminated have been thoroughly revised using a new set of methods for data validation and for the imputation of missing values. This note will help users become familiar with the new imputation approach and understand the differences that they may observe with respect to its previous release done in December 2015, through a set of revision indicators.

The revised agriculture production data cover the period 1991-2014. When compared to the previous release (for the period 1991-2013), data may have been revised due to:

- Update of official figures (new official data provided by national statistical institutions);
- Correction of errors discovered in official/semi-official figures;
- Replacement of previous imputed values with values produced with the new imputation strategy.

The new country data are being shared with each country for feedback and validation. Country validation is an additional feature of the new approach and will take place systematically before the annual database up-dates.

1. Background

Missing values are a common problem for international datasets, stemming from countries' non-responses to the data requests dispatched on a regular basis by international or regional organizations. Yet consistent non-sparse datasets are of critical importance to compile regional and global aggregates for the relevant indicators. The adoption of sound and robust imputation methods is therefore necessary to ensure that the assessment of regional and global trends for the key development outcomes is accurate and reliable.

Innovative methods have been developed by the FAO Statistics Division to improve data reliability and consistency across statistical domains. The new approach, in particular, employs all the available official information for imputing missing data. This allowed a backward revision of previous estimates and the use of information from similar countries to impute extensive data gaps. Significant research efforts have been placed on developing an innovative methodological approach to data imputation and solid software routines for the implementation of the new methods which ensures replicable statistical results.

The previous imputation methodology was mainly based on expert estimates, where available, and simple interpolations /extrapolations techniques. In addition the previous approach looked at each country-commodity combination separately while the new methodology embeds, though a mixed model, the ability to capture cross-country and cross-commodity information.

2. The Agricultural Production domain

FAO collects from national official sources area harvested (input) and agricultural production (output) data for each crop through a dedicated questionnaire dispatched on annual basis. In the case of meat products the questionnaire collects data on the number of slaughtered animals (input) and meat production (output). Data on agricultural productivity (e.g. yield for crops and carcass weight for meat), instead, are not collected. These three variables, however, are linked by the following relationship:

$$productivity = \frac{output}{input}$$

This relationship represents the key constraint for the imputation process and embeds the procedure to ensure that the elements of the triplet are always balanced.

3. Ensemble Imputation

The chosen approach to imputation is extremely flexible in order to capture the specificity of different country-commodity combinations and different farming practices.

To this end, the imputation method used in the Agricultural Production domain and other FAOSTAT domains is based on ensemble learning[1] on a set of models. This approach solves many of the shortcomings of the previous approach and offers a flexible and robust framework to incorporate additional information to further improve the performance of the estimates.

Ensemble learning refers to the process of building a collection of simple imputation models or learners which are later combined in a composite model to make a prediction. Ensembles are very popular in the data-mining community because of their ability to combine multiple models and produce an estimate that is more accurate and robust than any of the individual models.

The method consists of two steps:

1. **Building simple models/learners. The following ten models are used the Agricultural Production domain**
    a. Mean: Mean of all observations
    b. Median: Median of all observations
    c. Linear: Linear Regression
    d. Exponential: Exponential function
    e. Logistic: Logistic function
    f. Naive: Linear interpolation followed by last observation carried forward and first observation carried backward.

[1] See, for example, T. Dietterich: “Ensemble in Machine Learning”, available at: http://www.eecs.wsu.edu/~holder/courses/CptS570/fall07/papers/Dietterich00.pdf

g. ARIMA: Autoregressive Integrated Moving Average model selected based on the AICC, and imputation via Kalman Filter.
h. LOESS: Local regression with linear models and model window varying based on sample size.
i. Splines: Cubic spline interpolation.
j. MARS: Multivariate Adaptive Regression Spline
k. Mixed Model: Linear mixed model with time as a fixed effect and country and commodity as the random effects.

2. **Building combined predictions.** The ensemble imputation is constructed using a weighted average of all of the learners. However, models which perform poorly receive a much lower weight than models which fit the data better. This is accomplished using the cross-validation technique, where the observed data are split into different groups which are used to compare a model's predictions with the actual values. For each group, predictions are computed using all of the observed data, except for those in the group itself in order to measure how well this model estimates the data in the group. The average error across all groups, gives a measure of how well this model predicts the dataset. The error is then estimated for each of the ten models, to compute the model weights.

4. Revision indicators

The revisions introduced in the figures with the new data release are summarized by a set of revision indicators for the period 1991-2013. The ***Mean Revision*** (**MR**) is the average of the changes in a time series (new values vs. previous values). This indicator gives an assessment of the direction of changes, e.g. a negative average means that new values are, on average, lower than the previous ones. The impact of the changes (in absolute terms) is measured via the ***Mean Absolute Revision*** (**MAR**), which is usually expressed in relative terms (**RMAR**) to allow comparisons between changes in different items in the same country or over countries etc. RMAR is obtained as the ratio of MAR and the average of the final disseminated (imputed or existing) values in the time series.
Mean Revision is computed via the following expression:

$$MR = \frac{1}{n}\sum_{t=1}^{n}\left(X_{Lt} - X_{Pt}\right)$$

where

- X_{Pt} is "previously" released value of the given item at time reference *t*;
- X_{Lt} is the "latest" released value of the given item at time reference *t*;
- *n*: number of reference periods in the time series being considered.

The indicator applies to the cases where both X_{Lt} and X_{Pt} are not missing. If one or both of them are missing, then the couple of elements is discarded from the computation of the average.

MR provides indication on the direction of the revision but it does not provide information about the amount of the revisions. This can be measured by the ***Mean Absolute Revision***:

$$MAR = \frac{1}{n}\sum_{t=1}^{n}\left|X_{Lt} - X_{Pt}\right|$$

or the ***Relative Mean Absolute Revision****:*

$$RMAR = \frac{\sum_{t=1}^{n}\left|X_{Lt} - X_{Pt}\right|}{\sum_{t=1}^{n} X_{Lt}}$$

The RMAR is useful for comparative purposes as it is expressed in percentage terms.
These indicators where computed for the various food groups and regional groupings. The following sections summarize the major revisions occurred for crops, livestock primary and live animals.

5. Revisions for crops

The revisions of crops are greater than 10% for both area harvested (5312) and production quantity (5510) in a limited number of products and regions. The key items involved are ‘Jute & Jute-like fibres’ (1751), ‘Fibre Crops Primary’ (1753) and Vegetables (1735, 1800). Details can be found in the Tables 1a and 1b.

Table 1a – Average revisions of Area harvested (5312) by region (period 1991-2013). Cases with RMAR greater than 10%

AreaCode	AreaName	ItemCode	ItemName	MR	MAR	RMAR.L
5503	Micronesia	1814	Coarse Grain; Total	1.3	13.5	21.3%
5400	Europe	1751	Jute & Jute-like Fibres	1306.9	3916.3	21.3%
5401	Eastern Europe	1751	Jute & Jute-like Fibres	1306.9	3916.3	21.3%
5206	Caribbean	1751	Jute & Jute-like Fibres	528.7	598.3	18.7%
5803	Small Island Developing States	1751	Jute & Jute-like Fibres	528.7	598.3	18.7%
5504	Polynesia	1804	Citrus Fruit; Total	-226.9	255.8	18.6%
5206	Caribbean	1753	Fibre Crops Primary	6554.0	6554.0	16.5%
5803	Small Island Developing States	1753	Fibre Crops Primary	6503.8	6503.8	15.0%
5503	Micronesia	1717	Cereals;Total	-2.8	17.5	12.9%
5503	Micronesia	1817	Cereals (Rice Milled Eqv)	-2.8	17.5	12.9%
5204	Central America	1751	Jute & Jute-like Fibres	-472.3	472.3	12.3%
5105	Western Africa	1735	Vegetables Primary	-123882.5	253408.1	10.6%
5105	Western Africa	1800	Vegetables & Melons; Total	-123882.5	253408.1	10.6%

Table 1b – Average revisions of Production Quantity (5510) by region (period 1991-2013). Cases with RMAR greater than 10%

AreaCode	AreaName	ItemCode	ItemName	MR	MAR	RMAR.L
5206	Caribbean	1751	Jute & Jute-like Fibres	2449.8	2449.8	24.0%
5803	Small Island Developing States	1751	Jute & Jute-like Fibres	2449.8	2449.8	24.0%
5503	Micronesia	1814	Coarse Grain; Total	-1.0	18.6	19.1%
5104	Southern Africa	1751	Jute & Jute-like Fibres	155.2	167.2	16.8%
5206	Caribbean	1753	Fibre Crops Primary	4280.1	4704.0	16.6%
5803	Small Island Developing States	1753	Fibre Crops Primary	4241.3	4667.9	15.5%
5503	Micronesia	1841	Oilcakes Equivalent	-2132.6	2270.9	13.4%
5503	Micronesia	1732	Oilcrops Primary	-3465.3	3690.2	13.4%
5503	Micronesia	1804	Citrus Fruit;Total	-14.8	20.1	12.2%
5200	Americas	1751	Jute & Jute-like Fibres	4105.0	4448.5	11.3%
5102	Middle Africa	1729	Treenuts;Total	-150.8	164.5	10.9%
5400	Europe	1751	Jute & Jute-like Fibres	394.5	4948.2	10.7%
5401	Eastern Europe	1751	Jute & Jute-like Fibres	394.5	4948.2	10.7%
5504	Polynesia	1753	Fibre Crops Primary	-28.8	52.9	10.6%

6. Revisions for live animals

Revisions involving live animals stocks (5115 and 5112) are not significant as there are no cases with RMAR greater than 10% in any geographic area.

7. Revisions for primary livestock product

As far as ‘Milk, total’ (1780) is concerned, larger revisions for both ‘Milking Animals’ (5318) and ‘Yield’ (5420) are observed in Western Africa (5105) (see Table 3a).

Table 3a – Average revisions of Milking Animals (5318) and ‘Yield’ (5420) of ‘Milk, total’ (1780) by region (period 1991-2013). Cases with RMAR greater than 10%

AreaCode	AreaName	ItemCode	ItemName	ElementCode	MR	MAR	RMAR
5105	Western Africa	1780	Milk;Total	5318	-2,796,762.2	2,796,762.2	10.3%
5105	Western Africa	1780	Milk;Total	5420	181.6	181.6	15.5%

The cases with a RMAR greater than 10% for ‘Slaughtered Animals’ (5320 and 5321) are relatively few and concern ‘Sheep and Goat meat’ (1807) in Western Africa (5105) and Micronesia (5503); ‘Beef and Buffalo Meat’ (1806) in Micronesia (5503) (see Table 3b).

Table 3b – Average revisions of Slaughtered Animals (5320 and 5321) by region (period 1991-2013). Cases with RMAR greater than 10%

AreaCode	AreaName	ItemCode	ItemName	ElementCode	MR	MAR	RMAR
5503	Micronesia	1806	Beef and Buffalo Meat	5320	296.3	327.1	17.6%
5503	Micronesia	1807	Sheep and Goat Meat	5320	194.6	220.4	15.1%
5105	Western Africa	1807	Sheep and Goat Meat	5320	-3,896,790.1	6,308,474.9	11.7%

The same regions show larger revisions for ‘meat production’ (5510) (see Table 3c).

Table 3c – Average revisions of ‘meat production’ (5510) by region (period 1991-2013). Cases with RMAR greater than 10%

AreaCode	AreaName	ItemCode	ItemName	ElementCode	MR	MAR	RMAR
5503	Micronesia	1806	Beef and Buffalo Meat	5510	40.3	44.5	17.5%
5503	Micronesia	1807	Sheep and Goat Meat	5510	2.0	2.5	14.5%
5105	Western Africa	1807	Sheep and Goat Meat	5510	-59,197.6	73,934.8	11.6%